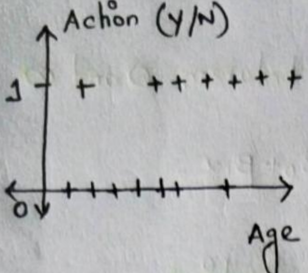
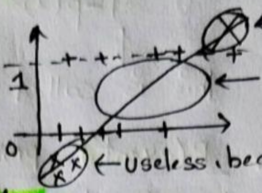
① LOGISTIC REGRESSION — 4 pages

Intution → Suppose we send a email to a set of customers, whether they respond to the email or not. (Action - Yes /No).



→ We can observe, more the age they respond to the email.

→ It is not linear regression model because.

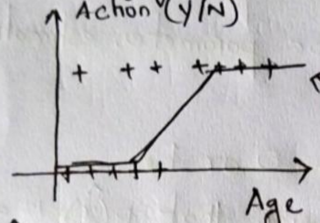


← Not good.

← Probability. If it is higher, then yes or else NO.

← useless, because probability can't be <0 or >1.

- We know probability 0-1.
- So, the target variable is either yes/No. So that is classification.



best fit line.

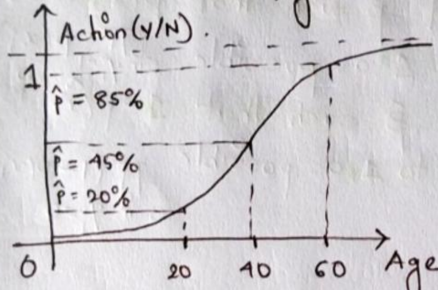
• Linear regression, $y = b_0 + b_1 * x$.

↓ Add sigmoid function.

$$p = \frac{1}{1+e^{-y}}$$

↓

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$ (logistic equation)



Mostly likely to respond - 60

Least likely to respond - 20.

if $\hat{p}$ is > 50 then yes for email.
$\hat{p}$ is =< 50 then No for email. [this number can change according to need].

Why logistic Regression?

→ Consider a dataset where response variable falls into one of two categories Yes or No. Rather than modelling the response Y directly, logistic regression models the probability that Y belong to particular category.

- For making a range of 0 to 1, we use logistic function, $p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1+e^{\beta_0 + \beta_1 x}}$

Method is known as ~~most~~ maximum likelihood function.

- Range is 0 to 1. Values cannot be less than 0 or more than 1.
- Logistic function will always produce S curve regardless of value of x.

$p(x) = \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$,

After bit of manipolation, $\dfrac{p(x)}{1-p(x)} = e^{\beta_0+\beta x}$.

Quontity, $\dfrac{p(x)}{1-p(x)}$ = odds, can take any value between 0 and ∞.

If we apply log then $\underbrace{\log\left(\dfrac{p(x)}{1-p(x)}\right)}_{\text{this is called logit function.}} = \beta_0+\beta_1 x$.

- "logit" = "log odds" ⇒ odd = $\dfrac{P(\text{event})}{1-p(\text{event})}$ $\dfrac{\text{probability event occur}}{\text{Probability event not occur}}$.
- Parameter Estimation -
  - Goal of learning is to estimate parameter vector $\hat{\beta}$.
  - Logistic regression uses Maximum likelihood for parameter estimation.
  - How does Maximum likelihood works ?
    - → Consider N samples with labels either 0 or 1.
    - → For sampled labelled "1": Estimated $\hat{\beta}$ such that $p(\hat{x})$ is as close to 1 as possible. $\prod_{s \text{ in } y_i=1} p(x_i)$
    - → For sample labelled "0": Estimate $\hat{\beta}$ such that $1-p(\hat{x})$ is as close to 1 as possible. $\prod_{s \text{ in } y_i=0} (1-p(x_i))$
- We have to optimise likelihood function.
- If data is not linearly seperable, we cannot apply logistic regression. Linearly seperable is property of two sets of points. Two sets of points are linearly seperable if there exists atleast one line in the plane which seperate the both set of points (blue and red).

Main steps in logistic Regression - Training set is given.

Step 1 - How to calculate logistic function. (to learn paramater of training set)

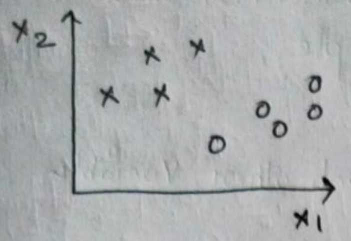
Step 2 - How to learn the coeffecient for a logistic regression model using stochastic gradiant

Step 3 - How to make prediction using the model.

Logistic f^n - logit f^n / sigmoid f^n to translate value of x to graph & learn the coefficient ($\beta_0/\beta_1/\beta_2$ etc). Build the model and predict.

Subalalitha Navaneethakrishnan

Let's considar



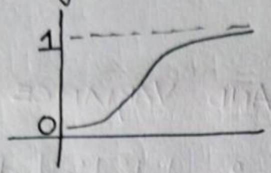
Step 1a- Linear seperable.

Step 1b - We need to transform data point using logit fn or Sigmoid f^n which is given by

$$\underset{(\text{logit } f^n)}{h_\theta(x)} = \frac{1}{1+e^{-x}} \text{ — } ①$$

Function transform each input value in range 0 to 1.

Step 2a After transformation, -ve number resulted in valve close to zero. The larger +ve number resulted in value close to one.



$$P(\text{class} = 0) = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2)}}$$

class = 0 or 1. modified logit f^n.

If Probability ≥ 0.5 then 1 or else 0.

Step 2b - We need to find paramter of eqn (b_0, b_1, b_2) with the help of gradient descent (to find best value of parameter).

Gradient descent → ① Suppose you are in top of mountain. & choose the best route to reach ground level.

Ⓘⓘ Take a step in any direction.

ⓘⓘⓘ Calculate distance at all angle and choose the smallest distance angle and take step.

ⓘⓥ Repeat step ⓘⓘⓘ till you reach the ground.

so, for first step choose 0 for parameter.

$$\text{prediction} = \frac{1}{1+e^{-(0+0(2.7)+0(2.3))}}$$

$$= 0.5$$

Ultimate we have to get best value (so that it will be equal to predicted) In short accuracy & minimum error.

- α is a learning rate, at which model learn.

Q) What is maximum likelihood function?
- Goal of ~~the~~ Maximum likelihood estimation is to make inferences about the population that is most likely to generate a sample.

Q) Assumptions in logistic regression?
① LR requires observation to be independent of each other. Variables should not be highly correlated.
(ii) Little or no multicollinearity among independent variables
(iii) LR assum linearity of independent variable and log odds.
(iv) LR typically require large sample size.

~~Topic~~ [illegible]

INTERPRETATION OF STANDARD DEVIATION AND VARIANCE

- Standard deviation $= \sqrt{\text{Variance}}$

→ Basically a small SD means that the values in statistical dataset are ~~means~~
~~the~~ close to the mean of the dataset, on average and a large standard
deviation means that the value in the dataset are farther away from
the mean, on average.

→ In short, it measure how concentrated the data around mean, more concentrated → smaller SD.

→ A small SD can be the goal in certain situations where result are restricted
for example in product manufacturing and quality control. A particular
type of car part that has to be 2 centimeters in diameter to fit
properly had better not to have a very big standard deviation during
manufacturing process. A big standard deviation would mean it will
end up in trash can because they don't fight right.

→ High SD, reflects a large amount of variation in the group. For example, if we look at the salaries for everyone in a company, including student intern to the CEO, standard deviation may be very large. On other hand, if we observe only student interns salary, standard deviation may be low/smaller.

→ Outlier does affect the SD, because formula includes the mean. SD cannot be negative and lowest possible value is 0 and 0 is possible only when every single entity have same number (no deviation)

→ SD have the same unit as the original data.